DOCUMENT RESUME

ED 183 594 TM 010 113

AUTHOR	Mills, Craig N.; Hambleton, Ronald K.
TITLE	Issues and Methods of Reporting Criterion-Referenced Test Scores. Laboratory of Psychometric and Evaluative Research Report No. 100.
INSTITUTION	Massachusetts Univ., Amherst. School of Education.
PUB DATE	[Oct 79]
NOTE	45p.; Paper presented at the Annual Meeting of the Northeastern Educational Research Association (Ellenville, NY, 1979)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Administrative Personnel; *Criterion Referenced Tests; Elementary Secondary Education; Guidelines; *Information Needs; *Information Utilization; Norm Referenced Tests; Parent School Relationship; *Scores; Student School Relationship; Student Testing; Teachers; *Test Interpretation; Test Results
IDENTIFIERS	Computer Test Scoring; *Test Reporting

ABSTRACT

The literature related to reporting, interpreting, and utilizing test results for norm-referenced and criterion-referenced testing is reviewed; and the prerequisites and other important characteristics of a good report are described. Guidelines are also provided for the development or evaluation of reporting systems for criterion-referenced tests or test batteries that combine criterion-referenced and norm-referenced tests. The following information is covered: the uses of scores; ways of reporting scores; the problems caused by limited knowledge of tests by score recipients, particularly students and their parents; difficulties in score interpretation; and effects of the use of computer technology. Prerequisites are given for appropriate reporting systems for criterion-referenced tests: identifying the audience for the reports and their interests; matching the program to those interests; matching the reports to the sophistication of those audiences; and choosing or constructing appropriate tests. A list of audiences, their information needs and uses are presented in tabular form. Additional recommendations are made concerning report format, the use of normative information, the value of flexibility in setting cutting scores, and problems of generalizability. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Craig N. Mills
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED183594

TM010 113

Issues and Methods of Reporting Criterion-Referenced Test Scores[1,2]

Craig N. Mills and Ronald K. Hambleton
University of Massachusetts, Amherst

Research during the past several years has led to the development of methods for the preparation and validation of criterion-referenced tests (Hambleton & Eignor, 1979; Millman, 1974; Popham, 1978). On the other hand, very little attention has been paid to the reporting and interpreting of the scores of the tests. For example, in two recent reviews of the criterion-referenced testing field (Hambleton, Swaminathan, Algina, & Coulson, 1978; Popham, 1978a) only a few sentences were devoted to the topics. The likely explanation is that measurement specialists have spent their time researching topics which precede logically the reporting and interpreting of test scores (for example, sorting out definitional problems, preparing methods for assessing content validity, assessing test score reliability, and determining test lengths).

It is unfortunate, however, that reporting and interpreting test scores have not received more attention. The purpose of a testing program is, after all, to provide usable information in a convenient format. Test score information that is inappropriate, confusing, or in any other way unsuited to the needs of potential test score users will be of limited value.

---

[1]Laboratory of Psychometric and Evaluative Research Report No. 100. Amherst, MA: School of Education, University of Massachusetts, 1979.

[2]A paper presented at the annual meeting of the Northeastern Educational Research Association, Ellenville, NY, 1979.

The quality and appropriateness of criterion-referenced test score reporting impacts directly on the extent of the use of test score information. Presently, there are millions of students taking criterion-referenced tests and they are at all levels of education. The decisions made from the results of the tests range from diagnosis of learning deficiencies and monitoring student progress in objectives-based programs to program evaluation and funding decisions. Many of these decisions have potential long-term implications for examinees. It is imperative, therefore, that the information provided to decision-makers be the appropriate type of information and that it be in a format which facilitates effective decision-making.

One might be tempted to suggest that reporting forms developed over the years for use with norm-referenced test scores with minor revisions could suffice. However, two reasons exist to explain the inappropriateness of using norm-referenced test score reporting practices. First, as will be discussed later, there are a large number of problems associated with current methods for reporting norm-referenced test score information. For example, small differences between scores are often over-emphasized by test users even when confidence bands of performance are reported. Second, the nature of the statements to be made about examinees is fundamentally different with criterion-referenced tests. Norm-referenced tests are constructed, principally, to facilitate comparisons among individuals (or groups) in relation to the performance of a norm group. Criterion-referenced tests, on the other hand, are developed to facilitate the interpretation of individual (or group) test performance in relation to a set

of objectives or competencies (Hambleton & Eignor, 1979). It is hardly surprising that approaches to reporting and using test scores will differ considerably since the primary purpose of criterion-referenced tests is different from norm-referenced tests.

The areas of criterion-referenced test score reporting and utilization require study since (1) little direct research has been carried out, (2) norm-referenced test score reporting technology is of limited value, and (3) the use of criterion-referenced tests has reached major proportions. The purposes of this study, therefore, are to (1) review the literature related to reporting, interpreting and utilizing test results in the areas of criterion-referenced and norm-referenced testing, (2) determine the qualities of a good report, and (3) provide guidelines for use in the development and evaluation of reports. The focus of this paper will be on two types of reporting systems. We are interested in the systems that accompany:

- criterion-referenced tests,
- combinations of criterion-referenced and norm-referenced tests.

We will not concern ourselves in this paper with statewide testing programs or programs which are solely concerned with reporting group information (for example, the National Assessment of Educational Progress).

The remainder of this paper is divided into four sections. Necessary antecedents to the preparation of high quality reports are considered in one section. The other three sections correspond to the three purposes of the study.

Review of Literature

Since proper reporting and interpretation of scores is so important, test developers and practitioners need to have at their disposal standards by which they can develop and evaluate reports of test scores. Although general guidelines exist for interpreting and reporting norm-referenced test scores, a great deal of dissatisfaction can be found for the manner in which test scores are used and interpreted. Page (1977) has said, "On one general problem of testing we can be in fair agreement: A great gap exists between the expertise of test development and the amateurish use of test scores" (p. 8-9). Fisher (1978) notes that there is no shortage of trained personnel to handle technical problems in the areas of test development, administration, and scoring. But, according to Fisher, "The problems begin when these results are communicated to various audiences, and the problems get serious when someone attempts to assign meaning to the data" (p. 35). Lewis (1977) reports that the "United Parents Organization recommended that Boards of Education set policies requiring principals to present and interpret test results in a comprehensible way to parents" (p. 17-18). Popham (1978b), while praising some aspects of the reporting system of the California Assessment Program notes "a certain half-heartedness in the explanatory documents that accompany CAP results" (p. 20). Hagen (1977) has also expressed concern:

> Those of us who have devoted our professional lives to testing and evaluation need to pay much more attention to translating test scores into constructive actions. Many of us have been too concerned with the predictive validity of tests and have been too little concerned with what test

> scores mean in terms of behavior and constructive actions to be taken to facilitate the development of the individual. We need to work more closely with teachers and educators to determine what information they need in order to make education more effective and help them get this information in a form which is useful to them. (p. 167)

Clearly, problems exist in the translation of test scores into useful decisions. Both norm-referenced and criterion-referenced tests are criticized with respect to adequacy of the reporting systems. Six areas of concern with respect to providing test score information to interested parties are described in the educational literature. The areas are:

1. Uses of test scores.
2. Manner of reporting scores.
3. Limited testing knowledge among teachers, parents, and students.
4. Presentation of results to parents and students.
5. Test score interpretation difficulties.
6. Use of computer technology to report test scores.

In the remainder of this review of literature each of the six areas will be briefly considered.

Uses of Test Scores

Stetz (1978) lists five major uses of test results: prediction, diagnosis, research, program evaluation, and assessment of achievement. The stated purposes of a testing program will determine to a large extent the use of the scores. However, the audiences for which the scores are intended is also an important consideration. Individuals

requiring information are students and their families, teachers, and administrators. Although some information will be desired by all groups, each group has unique information needs.

Students and their families need test information for predictive and diagnostic applications as well as for assessment of achievement (Stetz, 1978). The predictive application of test results may be used to chose appropriate curricula and to make educational and vocational choices (Goslin, 1967; Kirby, Culp, & Kirby, 1973). The diagnostic use of test results to identify an individual's strengths and weaknesses and to develop strategies for improvement or remediation is often cited (Bradley, 1978; Goslin, 1967; Hagen, 1977). When test results are used for determining achievement, the focus may be on achievement during a school year (Goslin, 1967), relative achievement in several different subjects or areas, or performance in one subject or area over time (Gardner, 1977).

Teachers may use test results as aids in making decisions about students and in evaluating their instruction. Examples of decisions about students are decisions about grouping and placement (Wahlstrom, Danley & Raphael, 1977) and group and individual diagnoses (Rost, 1973). Evaluation of instruction includes teacher self-assessment (Wahlstrom, Regan & Jones, 1978), curricular reform (Rost, 1973; Wahlstron, Danley & Raphael, 1977) and the appropriateness of the difficulty of course objectives (Wahlstrom, Regan & Jones, 1978).

At the local level a variety of administrative uses of test results are possible. Test results can be used to compare the performance of a school or system to national norms (Gardner, 1977; Wahlstrom, Danley & Raphael, 1977). Comparison of schools within a system may help identify patterns of achievement over time (Gardner, 1977) or to identify problem schools which may need additional resource personnel (Wahlstrom, Danley, & Raphael, 1977). Other administrative uses of test results include program evaluation and curriculum development (Goslin, 1967; Lawson & Ward, 1976; Stetz, 1978; Wahlstrom, Danley & Raphael, 1977) and the evaluation of teacher effectiveness (Goslin, 1967).

Manner of Reporting Scores

The type of statement made from test results is dependent upon the strategy utilized for measuring achievement. Ahmann (1978) and Millman (1978) list three strategies: item-centered, objective-centered, and subtest-centered. Therefore, test score reports can be centered around items, objectives, or subtests.

In an item-centered approach, information is presented about performance on each item in a test. Such a strategy can be employed to provide information about group performance on specific skills (reflected by single test items). It would not usually be advisable, however, to make statements related to individual examinee performance on an objective on the basis of performance on a single test item due to the unreliability of information provided by a single test item.

Objective-centered reporting involves making a statement about an individual's or group's performance on an objective on the basis of several items which measure that objective. A test may include many objectives and have many items per objective. Multiple test items/objective increases the reliability of reported examinee objective scores, but decreases the breadth of coverage of a test, unless testing time is increased.

Subtest-centered reporting usually involves a small number of skills, but many items measuring each skill. The skills are typically not extensively defined. The scores derived on the various components of a typical norm-referenced test (mathematics concepts, problem-solving, and computations; vocabulary, reading comprehension) are examples of subtest-centered reports.

A fourth strategy also exists. Millman (1970) suggests that two lists of objectives may be useful. One list would be for teachers. This list would include specific objectives to be measured. The second list, which would be for parents, would include broader objectives. Millman is suggesting that teachers receive information about performance on each objective whereas parents could receive information about clusters of related objectives. For example, Millman suggests that teachers may need information on objectives such as "identification of coins and converting coins to equivalent amounts of other coin values." Parents, on the other hand, might be confused by data on a large number of small objectives. It would be better to provide them with information such as "understands the dollar value of money" (p. 227).

Limited Testing Knowledge

One reason that teachers often misuse or misinterpret test scores is because they are unfamiliar with the field of tests and measurements. Goslin (1967) found that less than 40 percent of all teachers have had more than one course in test and measurement techniques. Many teachers have had no exposure to test and measurement techniques in either formal classes or in-service training. Over 50 percent of all elementary and private secondary school teachers (in the 1960's at least) had no formal test and measurement training. It is not surprising therefore to find that many teachers do not understand or properly use test results. The call for in-service and pre-service training to upgrade teacher competencies in interpreting and utilizing test results is widespread (for example, Dunn, 1969; Fleming, 1971; Lewis, 1977; Rost, 1973).

In most cases, parents are also lacking in competence in interpreting and understanding test results. Wahlstrom, Danley, and Raphael (1977) found most educators reluctant to present raw results to parents due to the perception that parents are unable to properly interpret the results. It was felt that parents might place too much emphasis and meaning on the scores. Others who have expressed concern about over-emphasis of test scores include Anastasi (1971), Backman (1976), and Brown (1976).

Presentation of Results to Parents and Students

One topic that has received much attention is the presentation of results to students. Since test results contain information of great potential benefit it is important that students receive the results in the best way possible. The results affect not only the student's intellectual response, but also his or her emotional response.

There is general agreement that reports of test results should be presented in face-to-face interpretive interview sessions (Backman, 1976; Bradley, 1978; Kirby, Culp, & Kirby, 1973; Miller, 1977; Thorndike & Hagen, 1977). In some situations, it would seem that group interpretive sessions are as effective as individual sessions (Folds & Gazda, 1966; Lallas, 1956; Rubenstein, 1978; Wright, 1963). Walker (1965) found individual sessions and group sessions equally effective when students were asked to recall scores, but that individual sessions lead to more acceptance of scores by examinees. While group sessions may be useful for explaining the concept of error in scores or for other general explanatory purposes, the potential effects of the scores on the student and his/her parents requires individual interpretive sessions.

Anastasi (1971) indicates a preference for the use of broad levels of performance and qualitative descriptions over numerical scores. Further, scores "should be accompanied by interpretative explanations by a professionally trained person" (p. 56). Backman (1976) also recommends reporting bands instead of numerical scores to reduce the chance of overemphasizing small differences in scores.

Test results should be interpreted in light of other information available about the examinee. Thorndike and Hagen (1977) make the following statement about a test interpretation:

> It should be set in the frame of reference of the particular student. Test scores should be interpreted in terms of what is known about the student's aptitude and about his educational or vocational goals. It should be directed toward positive and constructive action. It should emphasize the assets in a test profile or it should be oriented toward remedial action when achievement falls below what aptitude would lead one to expect. (p. 578)

The incorporation of background information is considered to be important in reporting student grades as well as in the interpretation of the results of standardized tests (Performance printouts for parents, 1974). For example, a computer reporting system in Memphis allows teachers to include anectodal information about the students in their reports to parents. Teachers select, from a list of statements stored on a computer, those which apply to each child. The printouts of the statements are sent home after the report cards to provide parents with descriptive information about their child's performance and work habits. The quotation above suggests that a similar system might be desirable when test results are reported as well.

Bradley (1978), Miller (1977), and Kirby, Culp, and Kirby (1973) suggest that test results be discussed with the examinee in light of his/her feelings on the day of the test and with reference to personal characteristics. Bradley and Kirby, et al., suggest that the quality of a test score is enhanced when the examinee may see some or all of

the test items. In many cases, however, it is not desirable to release actual test items. Popham (1978a, 1978b) recommends that the detailed statements of objectives measured by a test (called "domain specifications") be available upon request, but that they not be widely distributed. The length and detail of domain specifications render them too complex to be of use to most individuals. Popham recommends the use of "descriptive abstracts" which draw on those aspects of the test specifications which are directly relevant to instruction.

Test Score Interpretation Difficulties

Five major uses of tests and the corresponding types of tests which are most useful are listed in Table 1. It is clear from the table that many testing programs will require both norm-referenced and criterion-referenced interpretations. If this is the case, several options are available. It is possible to use a norm-referenced test with both norm-referenced and criterion-referenced interpretations. Most norm-referenced tests, however, do not have objectives or domains stated with sufficient specificity to allow for "strong" criterion-referenced interpretations (i.e., inferences cannot be made safely from examinee performance on a set of test items measuring an objective to a large class of behaviors defined by the objective). Also, the manner in which items are selected for inclusion in a norm-referenced test (deletion of items which are too easy or too difficult for an examinee group) does not facilitate criterion-referenced interpretations. A second option is to use a criterion-referenced test with both norm-referenced and criterion-referenced interpretations. The problem here is that norms

Table 1

The Major Uses of Tests, Their Purposes, and the Appropriate Type of Test Needed to Accomplish Each Purpose

Use of Test	Purpose	Type of Test Needed
Prediction	Differentiate among individuals on the basis of an ability or a trait.	Norm-referenced
Diagnosis	Determine what a particular individual can and cannot do.	Criterion-referenced
Research	Determine the relationship among variables	Norm-referenced
	Compare performance in experimental and non-experimental groups on well-defined tasks.	Criterion-referenced
Evaluation	Determine extent to which instruction has been effective in reaching program goals.	Criterion-referenced
	Determine achievement relative to that of other programs.	
	a. on program objectives	Criterion-referenced
	b. on global measures	Norm-referenced
Assessment of Achievement	Determine competence of students after instruction.	Criterion-referenced
	Determine relative achievement of individuals after instruction.	Norm-referenced

for criterion-referenced tests tend to be somewhat unstable for individual interpretations. Another alternative is to administer both a criterion-referenced test and a norm-referenced test. This approach may take more time and money than the others. On the other hand, the combined quality of NRIs and CRIs is apt to be better. A fourth possibility is the use of a single test battery that has a norm-referenced component and a criterion-referenced component. Such a battery allows for the best in criterion-referenced test development to be used in one section and the best in norm-referenced development to be used in the other. Users need not make psychometric sacrifices in order to obtain both criterion-referenced and norm-referenced interpretations. The primary advantages of approach four over approach three are the consistency of format and approach in the two tests and the availability of data (usually) on the inter-relationship of scores from the two tests.

Use of Computer Technology

The growth of computer technology has simplified the task of scoring tests and preparing reports. Baker (1971) points out that the speed and accuracy of the computer allows the use of various scoring keys with a single test to provide analyses beyond those usually provided. He also advocates the use of detailed verbal descriptions of student performance. Lewis (1977) has said that computer scoring allows publishers to provide much more information than is currently provided and to provide it in such a way as to help teachers diagnose learning problems.

Nichols and Knopf (1977) have listed several advantages of computerized score interpretations beyond those mentioned above. Such systems are faster and less expensive than individual systems. They can be used by persons not trained in test interpretation and are less subject to misinterpretation when read by different people than raw or standard scores.

Most available computer scoring systems do not, however, take full advantage of the available options. Furlong and Miller (1978) have pointed out that many scoring programs only provide students with reports of items missed and identification of the correct response for those items. Such reports do not provide information about an individual's performance relative to course objectives. Furlong and Miller (1978) describe a computer scoring program which provides individual reports of (1) an individual's performance relative to other students taking the test, (2) incorrectly answered items and correct responses to those items, (3) the objectives to which incorrectly answered items refer, and (4) if the instructor desires, alternative materials which may be used for further study. The program allows instructors to receive summaries of performance by item and by objective. The report also summarizes performance by taxonomic level of objectives.

It is clear that computer-generated reports and interpretations offer a promising, but as yet unfilled, alternative to traditional reports. They allow for a variety of scoring schemes, matching of report formats to audience needs and reduction of score misinterpretation.

Prerequisites for Appropriate Criterion-Referenced Test Score Reporting Systems

A test score report which provides needed information to several groups of interested parties is the product of much work. Several activities must occur before high quality reports can be prepared. The activities are listed below:

- Specification of Information Needs,
- Building a Testing Program Consistent with Needs,
- Identification of Audiences and Levels of Sophistication,
- Proper Test Selection,
- Proper Test construction.

The purpose of this section is to briefly discuss these necessary antecedents to the preparation of appropriate reports.

Specification of Information Needs

A school system should clearly specify the groups who are to be served by a testing program and what (specifically) their information needs are. From there, it is possible for a school district to formally state its purposes. Along the way, school subjects, course objectives, and grade levels which will be involved in testing should be specified.

Building a Testing Program Consistent with Needs

Two points are of concern. A test characteristic mentioned by Popham (1978a) is an adequate number of items per measured behavior. Although the number of items desired is not usually specified, a general idea of the relative emphasis to be placed on each domain of content

should be available. The other consideration is the scope of the program. A good testing program should provide data in relation to (at least) the most important information needs. It is important, however, that each testing situation include enough items per objective to yield reliable measurement of the objectives of interest at that time without requiring excessive amounts of testing time. It is necessary, therefore, to design a testing program which provides reliable and valid data on objectives and which presents the data in a manner which will not confuse the audiences or overload them with information which is either too specific or too general for their purposes.

Identification of Audiences and Levels of Sophistication

Decisions must be made about the types of information needed and the people who will receive the information. The lack of sophistication of teachers and parents in the field of tests and measurements has been discussed previously. This information should be considered when a school system determines the manner in which information will be presented to the various audiences. Specific statements of reporting goals at this stage can ease the burden of test selection and dissemination of results.

Proper Test Selection

If the purposes of the program have been adequately clarified as outlined above, test selection is considerably easier. The task is to identify the available tests which come closest to matching:

-the curricular emphases,

—the scope and focus, and

—the informational requirements of the program.

The task is not to identify the one test that exactly matches the program specifications. Such a search would in all probability be fruitless. The task is to identify a number of tests which come close to meeting the exact requirements. When the process of test selection is undertaken, one of the available guidelines for selection will be beneficial (Hambleton & Eignor, 1978; APA, 1974).

Proper Test Construction

One essential consideration is that of test development. A system which is considering purchasing a commercial test will find the previously mentioned selection guidelines helpful in assessing a test (Hambleton & Eignor, 1978). Some systems will want to develop their own criterion-referenced tests. Such a situation necessitates the availability of staff with both test development training and the time to do the job (Hambleton & Eignor, 1979).

Summary

The purpose of this section has been to emphasize the importance of several prerequisites of appropriate reporting systems. Although the fulfillment of these prerequisites will not guarantee the preparation of a high quality report, the failure to meet them will almost certainly insure low quality reports due to inappropriate or inaccurate data. The prerequisites are therefore necessary but not sufficient to insure the preparation of high quality reports of test results. The characteristics of reporting systems which do meet the needs of the several audiences interested in receiving test results and interpretations are considered in the next section.

Characteristics of Appropriate Test Score Reporting Systems

In this section elements and options are discussed which should be available in reports of criterion-referenced test scores. A logical analysis of the potential of criterion-referenced tests, current uses of the tests, and information needs of various audiences were used to generate recommendations reported in this section. Four audiences are addressed:

—Teachers,

—Parents and Students,

—Building Administrators, and

—District Administrators.

Table 2 provides a listing of the four audiences, the information to be reported, and the rationale for providing the information. Several of the information needs require explanation beyond that provided in the table. These needs are noted with a numerical superscript. Explanations are found in Appendix A. In a final section, four important characteristics which apply to all reports will be discussed.

Table 2

Audiences Desiring Test Results, Their Information Needs and Examples of Uses of the Information Provided

Audience	Information Needs	Use of the Information
Teachers	Master list of objectives tested.	Provide general comparison of test and curricular match
	Information keying items to objectives and objectives to clusters.	Provide specific comparisons of class activities and test.
	Individual student data by objective including raw scores and cut-off scores.	Identification of specific individual deficiencies and the degree of remediation necessary
	Individual student data by objective cluster.	Identification of general areas of individual deficiencies.
	Diagnostic statements of errors of non-masters.	Aid the design of instructional activities to upgrade performance.
	Performance of individuals on previous tests of the same or related objectives.	Identify trends in individual strengths and weaknesses.
	Identification by objective of all students who were classified as masters and those classified as non-masters.	Devise grouping patterns for new instruction and/or remediation.
	Summary class data for each objective.	Identification of specific instructional and/or curricular deficiencies
	Identification of objectives on which performance of the class was low.	Self-evaluation of instruction and determination of needs for group remediation.
	Summary class data for each cluster of objectives.	Identification of general areas of instructional and/or curricular deficiencies.

Audience	Information Needs	Use of the Information
	Previous performance of the class on the same or related objectives.	Identify trends in class performance.
	Previous performance of students in classes taught by this teacher on the same or related objectives.	Identify trends in effectiveness of curriculum and/or instruction.
	Performance of other classes at the same instructional level in the system.	Determine performance of the class relative to performance in the system.
	Performance of other classes at the same instructional level in the state or nation (optional).	Determine performance of the class relative to state or national performance.
Parents and Students[1]	Performance on clusters of objectives.	Provide general overview of performance.
	Identification of specific objectives on which performance is low.	Determine specific deficiencies.
	Inclusion of sample items from non-mastered objectives.	Clarification of skills to be mastered.
	Identification of trends of performance across tests or subtests.	Identification of strengths and weaknesse in broad areas of performance.
	Performance from previous tests on the same or related objectives.	Identification of trends of improvement or decline.
	Performance relative to other students in the same class.	Determine relative standing in the class.
	Performance relative to other students at the same instructional level in the system.	Determine relative standing in the system.
	Performance relative to other students at the same instructional level in the state or nation (optional).	Determine relative performance as compared to a national sample.
	Performance in relation to aptitudes.	Determine if student is performing to his/her potential.

Audience	Information Needs	Use of the Information
	Narrative diagnostic and interpretive reports to supplement numerical summaries.	Reduce misunderstanding of scores, provide alternative views of the data, identify areas needing attention.
	Statement from an official of the school system (if desired by system).[2]	Explain some aspect of the testing program.
	Comments from student's teacher.[3]	Provide background information to enhance interpretation of scores.
Building Administrators	Summaries of subtest performance for each classroom	Identification of classes which may need specific remediation.
	Summaries of subject performance for each classroom.	Identification of classes which may need general remediation. Determine the need for added personnel or in-service in a subject.
	Summaries of subtest performance by grade for the school.	Identify trends of performance.
	Identification, for each subtest, of clusters of objectives on which performance was low.	Indicate the need for curricular and/or instructional revision.
	Summaries of subject performance by grade for the school.	Identify subjects in need of curricular and/or instructional revision or increased resources.
	Summaries of subject performance by grade on previous tests of the same objectives.	Identify areas of improvement of decline.
	Summaries of student performance on key objectives.[4]	Monitor progress on school or district priorities.
	Summaries of student performance by grade for other district schools.	Comparison and identification of specific strengths and weaknesses. Identification of trends in the district.

Audience	Information Needs	Use of the Information
	Summaries of subject performance by grade for other district schools.	Comparison with other schools on a general basis. Identification of general performance trends.
	Master list of objectives and percentage of students classified as masters in each school and the district.	Reference and comparison.
	Individual permanent record labels.	Student files.
	Performance by grade relative to state or national norms for each subtest (optional).	Determine relative standing of classes of students.
District Administrators	Summaries of subtest performance in each grade by school.	Determine achievement levels.
	Summaries of subject performance in each grade by school.	Determine schools in need of additional resources (financial or special personnel).
	Summaries of subtest performance in each grade for the district.	Public release.
	Summaries of subject performance in each grade for the district.	Public release.
	Identification, for each subtest, those schools in which performance was low.	Determine in-service needs.
	Summaries of subject performance by grade on previous tests of the same objectives.	Identify trends of improvement or decline in the district.
	Master list of objectives, number of items per objective, cut-off scores, and percentage of students in the district exceeding the cut-off score.	Reference.
	Summaries of student performance on key objectives.[4]	Monitor progress on school or district priorities.

Audience	Information Needs	Use of the Information
	"Split" summaries of subject performance by designated subgroups (race, sex, etc.).	Public release, reports to government officials.
	Normative data of subtest performance relative to the state or nation.	Comparison of achievement with other districts.
	Computer tapes containing "raw" data of student performance.	Research studies within the system.

Important Characteristics for All Reports

There are a number of characteristics of report forms which are important, regardless of audience. They are:

1. Physical considerations
2. Reporting normative information
3. Flexibility of cut-off scores
4. Generalizability of the test scores.

Each of the characteristics will be considered next.

Physical Considerations. The size of the report can be a problem. Small reports are easy to lose or damage. Large reports are cumbersome and hard to store in standard folders or notebooks. Therefore, reports should be printed on standard 8½" x 11" paper. Each page should list the audience to receive the report, a date, the information included on that page, and the examinee or group of examinees covered by the information. Reports should arrive in the format in which they will be distributed. That is, school personnel should not be required to fold, cut or paste reports for the different audiences.

Whenever possible, all information pertaining to one test or subtest should be included on a single page. This eliminates the need for referring back and forth between pages to make comparisons. Attempts to provide all data on one page should not, however, forsake legibility; sufficient space should be allowed between columns and rows of scores to allow easy reading. Reports which have alternating rows or columns of shaded and nonshaded background facilitate legibility. Narrative passages should be within a page of the tables to which they refer if they cannot be included on the same page. Not only does this keep related information together, but it also separates tables of numbers from one another which improves the ease of reading the report.

Reporting Normative Information. If the tests include a norm-referenced component, the norm-referenced information for a test should be included with the criterion-referenced information for the same test. Norm-referenced information for all tests or subtests should not be grouped together on a separate sheet. To do so invites confusion and

misinterpretation. Norm-referenced interpretations should always be reported as bands of numbers or as numbers including error terms. (Eightieth to ninetieth percentile or eighty-fifty percentile $\pm$ five percentiles.)

Flexibility of Cut-off Scores. One quality that greatly enchances the value of the reports is allowing school systems to choose the cut-off score for each objective. Since schools place different importance on different objectives it is reasonable to assume that students would be expected to perform better on some objectives than others. School systems could receive instructions on procedures which could be used to choose an appropriate cut-off score for each objective. Alternately, it is possible to provide school systems with a list of objectives and three possible cut-off scores for each objective which could be chosen to reflect the level of importance placed on the domain at a certain grade level in the system.

Generalizability of the Test Scores. Many of the tests which are currently called criterion-referenced tests are more accurately described by the term objective-referenced tests. The difference is an important one. An objective-referenced test is one in which items are keyed to behavioral objectives. The scores on such a test reflect an examinee's ability on those items which make up the test. A criterion-referenced test, on the other hand, is composed of items which represent a sample from a well-defined content or behavior domain. Such a test allows an examinee's score to be interpreted not only in relation to the items on the test, but also in relation to the entire domain of behavior sampled by the test. It is the latter interpretation that is most often desired (so much so, that often such interpretations are

made even when the domain of behavior has not been specified). A test score report should include a section describing the generalizability of the test scores. Failure to provide such information invites over- or under-interpretation of the scores.

Summary. The elements which should be found in reports of test scores have been briefly considered. Four different audiences were considered: teachers, parents and students, building administrators, and higher level administrators. Each audience has different needs and should receive reports which address those needs. Several characteristics were discussed which apply to all reports. These include physical considerations, placement of norm-referenced information, flexibility of cut-off scores and generalizability of the test scores.

Guidelines for Evaluating Score Reporting Systems

This section of the paper provides questions which can be used to evaluate or guide the development of reports of criterion-referenced test results. The questions are broken into six sections reflecting:

1. Audiences to whom reports should be provided.
2. Components of teacher's reports.
3. Components of reports received by parents.
4. Components of reports received by building administrators.
5. Components of reports received by higher level administrators.
6. General considerations for all reports.

All questions are worded positively, that is, if the report is in line with recommendations of the previous section, the answer to the question would be yes. It is suggested, however, that yes-no responses not be used. Instead, answers should be "S", "E", or "N". "S" would indicate that the information is provided as part of the standard reporting package of the test. "E" indicates that the information can be provided, but that an extra charge is involved. "N" is used to indicate that the information or service is not available.

1. Audiences

1.1 Are reports available for classroom teachers? _____

1.2 Are individual student reports available for students and their parents? _____

1.3 Are reports available for building administrators? _____

1.4 Are reports available for higher level administrators such as superintendents and their assistants? _____

2. Teacher Reports

2.1 Are all objectives or (domains) measured by the test listed? ____

2.2 Are the items which represent each domain identified? ____

2.3 Is the total number of items measuring each objective clearly defined? ____

2.4 Is the cut-off score which was used to assign examinees to mastery states on each objective identified? ____

2.5 Is the raw score (or percent score) of each child on each domain printed? ____

2.6 Are students who have been classified as masters identified for each objective? ____

2.7 Is summary data on class performance available for each objective (average percent scores)? ____

2.8 Are clusters of related objectives identified? ____

2.9 Is performance of each student on each of the clusters provided? ____

2.10 Is summary data of class performance on each of the clusters provided? ____

2.11 Are individuals whose performance is sub-standard listed for each objective? ____

2.12 Are diagnostic statements available about the errors of each examinee? ____

2.13 Are objectives identified on which total class performance was relatively low? ____

2.14 Is information pertaining to individuals' previously identified strengths and weaknesses provided (after the first year)? ____

2.15 Is information pertaining to strengths and weaknesses identified in previous classes taught by the teacher provided (after the first year)? ____

2.16 Are summaries of performance of other classes at the same instructional level in the system available for each cluster of objectives? ____

2.17 Are summaries of performance of other classes at the same instructional level in the state or nation available (optional)? ____

3. Parent and Student Reports

3.1 Is performance reported for each cluster of objectives? ____

3.2 Within each cluster are the objectives on which performance was substandard identified? ____

3.3 Are example items included in the identification of objectives in which performance was substandard? ____

3.4 Are common sources of errors which occur across tests or subtests identified? ____

3.5 Are improvements or declines in performance from previous test administrations noted (after the first year)? ____

3.6 Is performance reported in relation to aptitudes? ____

3.7 Is the typical performance of other students in the same class identified for each cluster of objectives? ____

3.8 Is the typical performance of other students at the same instructional level in the system identified for each cluster of objectives? ____

3.9 If norms are reported are they reported as bands rather than specific percentile ranks? ____

3.10 Are diagnostic statements included which refer to objectives in which performance was low? ____

3.11 Is it possible for a standard statement from the superintendent (or another official) to be included in the report? ____

3.12 Is there a section of the report which includes comments about each child which teachers have chosen from a list of standard statements? ____

4. Building Administrator Reports

4.1 Are summaries of performance on each subtest available for each classroom? ____

4.2 Are summaries of performance on each subtest available by grade? ____

4.3 Are summaries of performance on each subtest available for other schools in the district? ____

4.4 Are summaries of subject performance available for each classroom? ____

4.5 Are summaries of subject performance available by grade? ____

4.6 Are summaries of subject performance available for other schools in the district? ____

4.7 Are summaries of past performance of each school provided for each subtest (after the first year)? ____

4.8 For each subtest, are clusters of objectives identified on which performance in the system was low? ____

4.9 Is a master table which identifies school performance on all objectives provided? ____

4.10 Are individual scores provided in a manner which facilitates placing them in permanent student record files? ____

4.11 Are summaries of student performance on key objectives available? ____

4.12 Are norms reported for use in judging the school against others in the state or nation? ____

5. Higher Level Administrator Reports

5.1 Are summaries of subtest performance available by grade for each school? ____

5.2 Are summaries of subject performance available by grade for each school? ____

5.3 Are district summaries of subtest performance available? ____

5.4 Are district summaries of subject performance available? ____

5.5 Are schools which perform poorly identified for each subtest? ____

5.6 Are results of previous tests of the same objectives available by grade for each subject? ____

5.7 Is a master list of objectives provided? ____

5.8 Is the number of items for each objective listed? ____

5.9 Are cut-off scores included? ____

5.10 Are summaries of percent masters in the district provided for each objective? ____

5.11 Is information provided which relates to student performance on designated key objectives? ____

5.12 Are "split" summaries of performance of designated subgroups available (by race, sex, etc.)? ____

5.13 Is normative data provided? ____

5.14 Is a computer tape of "raw" student data available? ____

6. General Considerations

6.1 Are all reports on 8½" x 11" paper? ____

6.2 Does each page of the report identify the audience to receive the report? ____

6.3 Does each page of the report identify the information on that page? ____

6.4 Does each page of the report identify the examinee or group of examinees for which information is provided on that page? ____

6.5 Is the test data clearly identified on each page of the report? ____

6.6 Is all information about one test or subtest included on the same page whenever possible? ____

6.7 Are rows and columns of numbers well spaced or placed on backgrounds of different shades to facilitate legibility? ____

6.8 Are narrative passages within one page of the numerical information to which they refer? ____

6.9 If norm-referenced information is reported, is the information included with relevant criterion-referenced information? ____

6.10 Is norm-referenced information always reported as a band or as a number with an error term provided? ____

6.11 Are systems able to chose a cut-off score for each objective in order to allow local curricular emphasis to influence mastery decisions? ____

6.12 Are reports provided in a form which does not require system personnel to further prepare the reports before dissemination (cutting, folding, pasting)? ____

6.13 Is a section of the report devoted to a discussion of the generalizability of the test scores? ____

Summary

It is clear that current reporting systems for use with criterion-referenced tests are not satisfactory. In this paper we have discussed the relevant literature and the qualities necessary in a high quality report. Also, we have provided a set of guidelines for reporting systems. At least two tasks lie ahead. First, using the guidelines presented here, examples of high quality reporting systems should be prepared. These reports would serve as references for others as they develop reporting systems to accompany criterion-referenced testing programs. Second, the guidelines presented in this paper should be used to evaluate many of the reporting systems which accompany currently available criterion-referenced tests. Such evaluations would be helpful for school systems as they consider the selection of a testing system to provide necessary information for effective decision making.

References

Ahmann, J. S. Basic issues concerning competency-based testing. In R. B. Ingle, M. R. Carroll, & W. J. Gephart (Eds.), Assessing student competence in the public schools. Bloomington, In: Phi Delta Kappa, 1978, pp. 78-89.

American Psychological Association. Standards for educational and psychological tests. Washington, D.C.: American Psychological Association, 1974.

Anastasi, A. Psychological testing (4th ed.). New York: MacMillan, 1976.

Backman, M. E. Reporting test results for guidance and instructional purposes. In Guide for School Testing Programs, National Council on Measurement and Evaluation, 1976.

Baker, F. B. Automation of test scoring, reporting, and analysis. In R. L. Thorndike (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1971, pp. 202-234.

Bradley, R. W. Person-referenced test interpretation: A learning process. Measurement and Evaluation in Guidance, 1978, 10, 201-210.

Brown, F. G. Principles of educational and psychological testing (2nd ed.). New York: Holt, Rinehart, and Winston, 1976.

Dunn, S. S. Helping teachers to make better use of test results. In K. Ingekamp (Ed.), Developments in educational testing. London: University of London Press, 1969.

Fisher, T. H. The assessment of student competencies in Florida's public schools. In R. B. Ingle, M. R., Carroll, & W. J. Gephart (Eds.), The assessment of student competence in the public schools. Bloomington, IN: Phi Delta Kappa, 1978, pp. 24-38.

Fleming, M. Standardized tests revisited. The School Counselor, 1971, 19, 71-72.

Folds, J. H., & Gazda, G. M. Comparison of the effectiveness and efficiency of three methods of test interpretation. Journal of Counseling Psychology, 1966, 13, 318-324.

Furlong, F., & Miller, W. DIAGNOSE: Computer-based reporting of criterion-referenced test results. Educational Technology, 1978, 18, 37-39.

Gardner, E. F. Interpreting achievement profiles: Uses and warnings. Journal of Research and Development in Education, 1977, 10, 51-63.

Goslin, D. A. Teachers and testing. New York: Russel Sage Foundation, 1967.

Hagen, E. Use and abuse of tests in education. In R. M. Bossone and M. Weiner (Eds.), Proceedings of the National Conference on Testing: Major Issues. University of New York. November 1977. (ERIC Document Reproduction Service No. ED 152 814.)

Hambleton, R. K., & Eignor, D. R. Guidelines for evaluating criterion-referenced tests and test manuals. Journal of Educational Measurement, 1978, 15, 321-327.

Hambleton, R. K., & Eignor, D. R. A practitioner's guide to criterion-referenced test development, validation, and test score usage. Laboratory of Psychometric and Evaluative Research Report No. 70 (2nd ed.). Amherst, MA: School of Education, University of Massachusetts, 1979.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Kirby, J. H., Culp, W. H., & Kirby, J. MUST: Manual for the users of standardized tests. Bensonville, IL: Scholastic Testing Service, 1973.

Lallas, J. E. A comparison of three methods of interpretation of the results of achievement tests to pupils. Dissertation Abstracts, 1956, 16, 1842.

Lawson, J. H., & Ward, A. W. Use of tests in program evaluation. In Guide for School Testing Programs. National Council on Measurement and Evaluation, 1976.

Lewis, B. Testing: A parent's point of view. In R. M. Bossone and M. Weiner (Eds.), Proceedings of the National Conference on Testing: Major Issues. University of New York, November 1977. (ERIC Document Reproduction Service No. ED 152 814.)

Miller, G. M. After the testing is over. Elementary School Guidance and Counseling, 1977, 12, 138-143.

Millman, J. Reporting student progress: A case for a criterion-referenced marking system. Phi Delta Kappan, 1970, 52, 226-230.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, CA: McCutchan Publishing Co., 1974.

Millman, J. Strategies for constructing criterion-referenced assessment instruments. Paper presented at the eighth annual conference on large-scale assessment, Boulder, Colorado, 1978.

Nichols, M. P., & Knopf, I. J. Refining computerized test interpretations: An in-depth approach. Journal of Personality Assessment, 1977, 41, 157-159.

Page, E. B. Testing: What in the world are we arguing about. In R. M. Bossone and M. Weiner (Eds.), Proceedings of the National Conference on Testing: Major Issues. University of New York. November 1977. (ERIC Document Reproduction Service No. ED 152 814.)

Performance printouts for parents. Nation's Schools and Colleges, September 1974, 1, 31-32.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall, 1978. (a)

Popham, W. J. Practical criterion-referenced measures for intrastate evaluation. Educational Technology, 1978, 18, 19-23. (b)

Rost, P. Useful interpretation of standardized tests. Clearing House, January 1973, 47, 319-320.

Rubenstein, M. R. Integrative interpretation of vocational interest inventory results. Journal of Counseling Psychology, July 1978, 25, 306-309.

Stetz, F. P. Providing relevant data for decision-making purposes. The Elementary School Journal, June 1978, 78, 220-225.

Thorndike, R. L., & Hagen, E. P. Measurement and evaluation in psychology and education (4th ed.). New York: John Wiley & Sons, 1977.

Wahlstrom, M. W., Danley, R. R., & Raphael, D. Measuring achievement at the primary and junior levels: An analytical review of test instruments used in evaluating pupil achievement and of communicating results to parents. Toronto, Ontario: Ontario Institute for Studies in Education, 1977.

Wahlstrom, M. W., Regan, E., & Jones, L. L. An analysis of teacher beliefs in relation to procedures for assessing student achievement. Toronto, Ontario: Ontario Institute for Studies in Education, 1978.

Walker, J. L. Four methods of interpreting test scores compared. Personnel and Guidance Journal, 1965, 44, 402-404.

Wright, E. W. A comparison of individual and multiple counseling for test interpretation interviews. Journal of Counseling Psychology, 1963, 10, 126-134.

Appendix A

Notes to Accompany Table 2

1. It should be noted that the data provided to parents is less specific than the data provided to teachers. While teachers need specific information on every objective in order to devise instructional prescriptions, parents and students may only require information about performance on clusters of related objectives. A report of an arithmetic computation test might include information about the student's overall performance in addition, multiplication, subtraction, and division. Further subdivisions such as sutraction of whole numbers, subtraction of decimals, subtraction of fractions, etc., might provide too many sets of data for the parents. It would be better to provide an overall performance appraisal for subtraction and then identify areas which need further work.

   For example, a child might answer 28 out of 35 subtraction problems correctly, but only correctly answer 3 out of 7 questions dealing with the subtraction of fractions. The report to the parent would say that the student had answered 80% of the items which related to subtraction correctly, but that subtraction of fractions was an area where performance was low. Statements of objectives and example items for those areas which show less than adequate achievement should be included.

2. Communication between schools and parents is often neglected. When reports of test results reach the parents they are often unaware of the purpose or scope of the testing program. A short statement from the superintendent or some other official would enhance the acceptance and understanding of the program and the scores reported.

3. Teachers are in possession of a wealth of information which could enhance student and parent understanding of test score reports. Teachers could receive a coded list of statements concerning classroom activities. Teachers could select, from the list, those statements which apply to each student. The codes could be recorded on the student's answer sheet. A computer program could then include the statements in the individual reports. Statements could range from identification of objectives which have not yet been taught to statements pertaining to an individual's interest in a given subject area.

4. Often a school or school district will choose a small number of key objectives on which to concentrate in a given year. The option should exist for a number of objectives (2-3 per subject area) to be classified as key objectives. Data on these key objectives should be presented to building administrators and to system administrators.